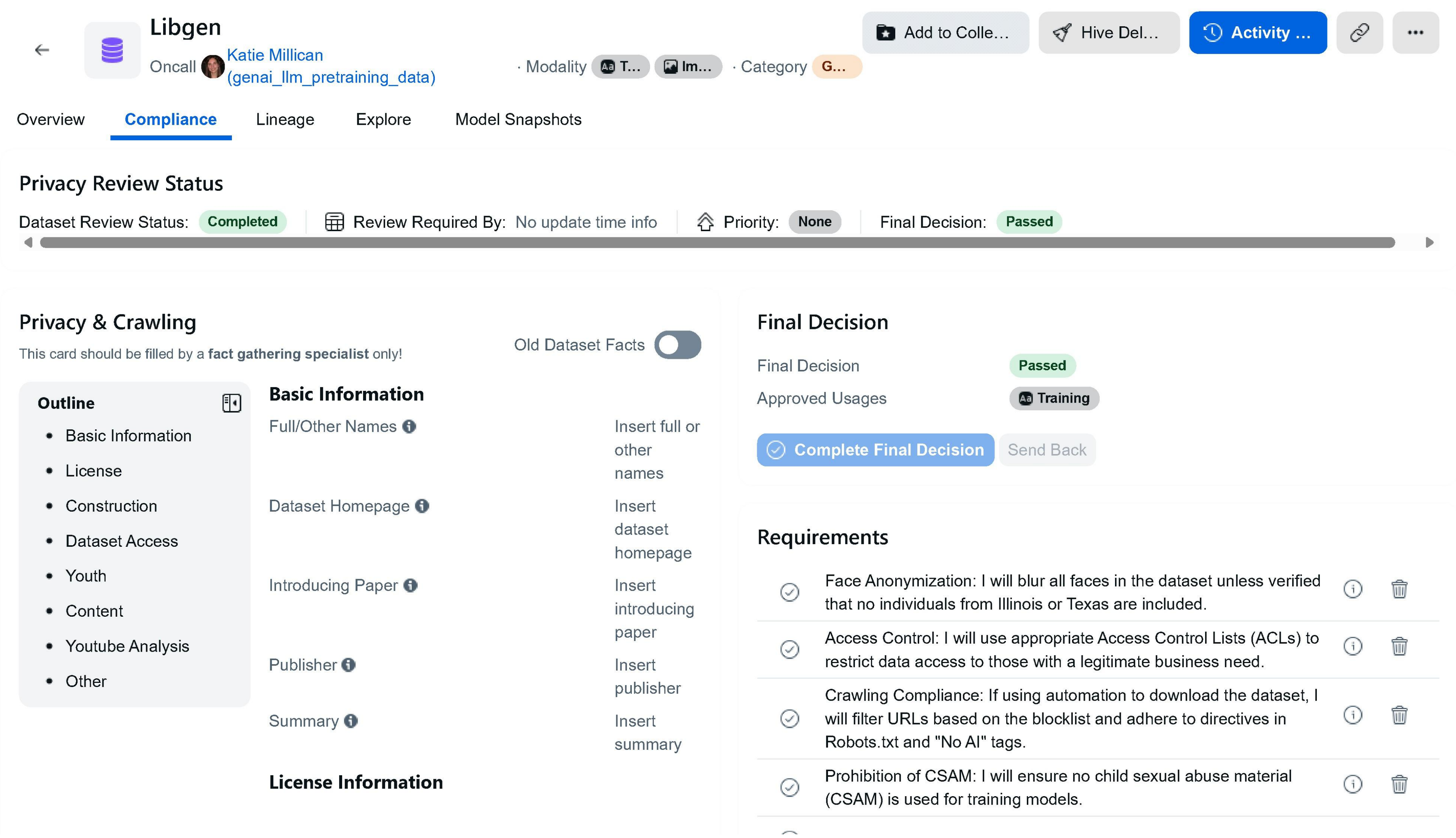
EXHIBIT K



Libgen

Oncall Katie Millican (genai_llm_pretraining_data) · Modality T… Im… · Category G…

Privacy Review Status

Dataset Review Status: Completed | Review Required By: No update time info | Priority: None | Final Decision: Passed

Privacy & Crawling

This card should be filled by a **fact gathering specialist** only!

Old Dataset Facts

Outline

- Basic Information
- License
- Construction
- Dataset Access
- Youth
- Content
- Youtube Analysis
- Other

Basic Information

Full/Other Names	Insert full or other names
Dataset Homepage	Insert dataset homepage
Introducing Paper	Insert introducing paper
Publisher	Insert publisher
Summary	Insert summary

License Information

Final Decision

Final Decision	Passed
Approved Usages	Training

Requirements

- Face Anonymization: I will blur all faces in the dataset unless verified that no individuals from Illinois or Texas are included.
- Access Control: I will use appropriate Access Control Lists (ACLs) to restrict data access to those with a legitimate business need.
- Crawling Compliance: If using automation to download the dataset, I will filter URLs based on the blocklist and adhere to directives in Robots.txt and "No AI" tags.
- Prohibition of CSAM: I will ensure no child sexual abuse material (CSAM) is used for training models.

License Types ℹ	Select license types
License Notes ℹ	Insert license notes

Construction Information

Construction Methods ℹ	Select construction methods
Data Origin Source ℹ	Insert data origin source
Domain Blocklist ℹ	Insert domain blocklist here
TV Shows, Movies, Stock Photos ℹ	Insert included media here
Construction Notes ℹ	Insert construction notes

If Crawled

Meta or 3P Crawled ℹ	Insert meta or 3P crawl here
Login Required? ℹ	Insert login required here

✓ Data Retention Limit: I will not retain data for longer than 1095 days.

✓ Handling of 1PD in 3PD Dataset: If I detect first-party data (1PD) in the 3PD dataset, I will cease its use and consult the Dataset Review Team.

Does Robots.txt Restrict Access? ⓘ	Insert robot restrict access here
Content or IDs/URLs? ⓘ	Insert crawl content type here

If Crawled or Combined from Existing:

Full Curation / Annotation? ⓘ	Insert full curation or annotation here
Content Fully from Trusted Sources? ⓘ	Insert is trusted source here

If Annotated / Derived / Combined from Existing:

Source Dataset(s) ⓘ	Insert source datasets

If Production:

Does this 3P dataset contain 1PD? ⓘ	Insert contains 1PD here

If Synthetic:

Does dataset contain synthetic data? ⓘ	Insert is contains synthetic data here

Dataset Access Information

Access Method ⓘ	Insert access

	method here
Access Notes	Insert access notes

Youth

Contains Youth data (u18)?	Choose Yes/No
Is Youth (u18) data the focus or incidental?	Insert data presence here

Text Content Information

Contains Text?	Choose Yes/No
Is there PII present in the dataset?	Choose Yes/No
If PII is present, is it incidental or the focus of the dataset?	Insert data presence here
Is News content present in the dataset?	Choose Yes/No
If News content is present, is it incidental or the focus of the dataset?	Insert data presence here
Text Content Notes	Insert text content notes

Visual Content Information

Contains Images or Video?	Choose Yes/No

Contains People in Images or Video?	Insert contains people here
Do we know provenance of images or video?	Insert provenance image or video here
IL/TX for People in Images or Video?	Insert image and video IL/TX here
Image or Video Annotations?	Insert images or videos annotation here
Visual Annotation Type(s)	Select visual annotation types
Image or Video Annotations Involve Humans?	Insert annotation involve human here
Visual Content Notes	Insert visual content notes

Audio Content Information

Contains Audio?	Choose Yes/No

Contains Music?	Insert contains music here
Contains Music Lyrics? (text)	Insert contains music lyrics here
Audio Annotations?	Insert contains audio annotations here
Audio Annotations Type	Insert audio annotations type here
Human Voices Present Within Audio?	Insert human voices presented here
IL/TX for Speaking People?	Insert is there IL/TX for Speaking People here
Audio Content Notes	Insert audio content notes

YouTube Analysis

Publisher Y/N	Insert youtube

	publisher
# Citations ℹ	Insert number of citations
Citations Link ℹ	Insert citations link
Top 15 Citation(s) ℹ	Insert top 15 citations
Citations ℹ	Insert is citations
Info ℹ	Insert google info
Is Google Crawled ℹ	Insert is google crawled
Industry Standard Final Determination ℹ	Insert industry standard final determination here
Industry Standard? ℹ	Insert industry standard
Chinese Origin? ℹ	Choose Yes/No
If Chinese Origin:	
Government Suppliers ℹ	Insert government

	suppliers here
Surveillance Data	Insert surveillance data here
Sensitive Topics and Subject Matter Domains	Insert sensitive topics and subject matter domains here
Is Present on EU/US Notorious Markets List or Piracy Lists	Choose Yes/No
File names include 'Pirated' or 'Stolen'	Choose Yes/No
Reputation for hosting or providing Pirated material	Choose Yes/No
Restricted Data Present	Insert restricted data presence here
CBRNE Content Present	Insert CBRNE content presence here

Other

Other notes	Insert other notes

Activity & Comments

Type a comment

Morphing Framework Bot added source: https://libgen.is/
February 17, 10:36 AM · Like

Morphing Framework Bot updated modalities to [text, image]
January 8, 6:42 AM · Like

Morphing Framework Bot created this AIDC Dataset.
December 12, 2024 · Like

Morphing Framework Bot made several changes
December 12, 2024 · Hide updates

- added the tag aidc_bespoke.
- added the tag genai.
- added the tag llama3_pretraining_data.
- added the tag wave 3.

Eyal Hochman Set the table 'scitech_en_union_train' from namespace 'gen_ai' as a seed
September 19, 2024 · Like

Lilach Mor made several changes

September 19, 2024 · Hide updates

- updated the "Model Family" to "LLAMA3, LLAMA4"
- updated the "Used Modality" to "Text"
- added the tag llama3_pretraining_data.

TDM Processor Completed the stage Requirements
September 16, 2024 · Like

Ramakant Shankar made several changes
July 23, 2024 · Hide updates

- added the tag aidc_bespoke.
- changed the description. · View Changes

Yoni Goyhman made several changes
June 16, 2024 · Hide updates

- updated the "Usages" to "Training"
- created this dataset